

本周周报(11.19-11.25):

解聪

本周工作:

时序异常交易分析

本周拿到了淘宝交易一周的真实交易数据。

数据

包含了 10 个类目: 3C 数码配件市场, 居家日用/收纳/礼品, 电子词典/电纸书/文化用品, 书籍/杂志/报纸, 饰品/流行首饰/时尚饰品新, 移动/联通/电信充值中心, 玩具/模型/动漫/早教/益智, 服饰配件/皮带/帽子/围巾, 零食/坚果/茶叶/特产, 童装/童鞋/亲子装。

时间范围为一周, 从 2011 年 9 月 19 日 0:00 到 9 月 25 日 23: 59。平均每天有近 400 万条交易, 总共有 2600 多万条交易记录。

交易数据的维度还是和先前的数据一样: date, buyer_id, seller_id, auction_price, buy_amount, aa_city, aa_prov, cat_id, cat_name, cat1, name1, lgo_prov, lgo_city。依次代表: 交易时间, 买家 ID, 卖家 ID, 单笔交易价格, 单笔交易数量, 卖家城市, 卖家省份, 叶子类目 ID, 叶子类目名称, 大类目 ID, 大类目名称, 买家城市, 买家省份。

统计后得到一周中这八个类目的交易总价格以及总量分布如下:

从图中可以发现一周交易具有明显的周期性。交易金额与总量的分布以一天为周期, 在每天的午饭前后以及晚饭后出现短暂的峰值。

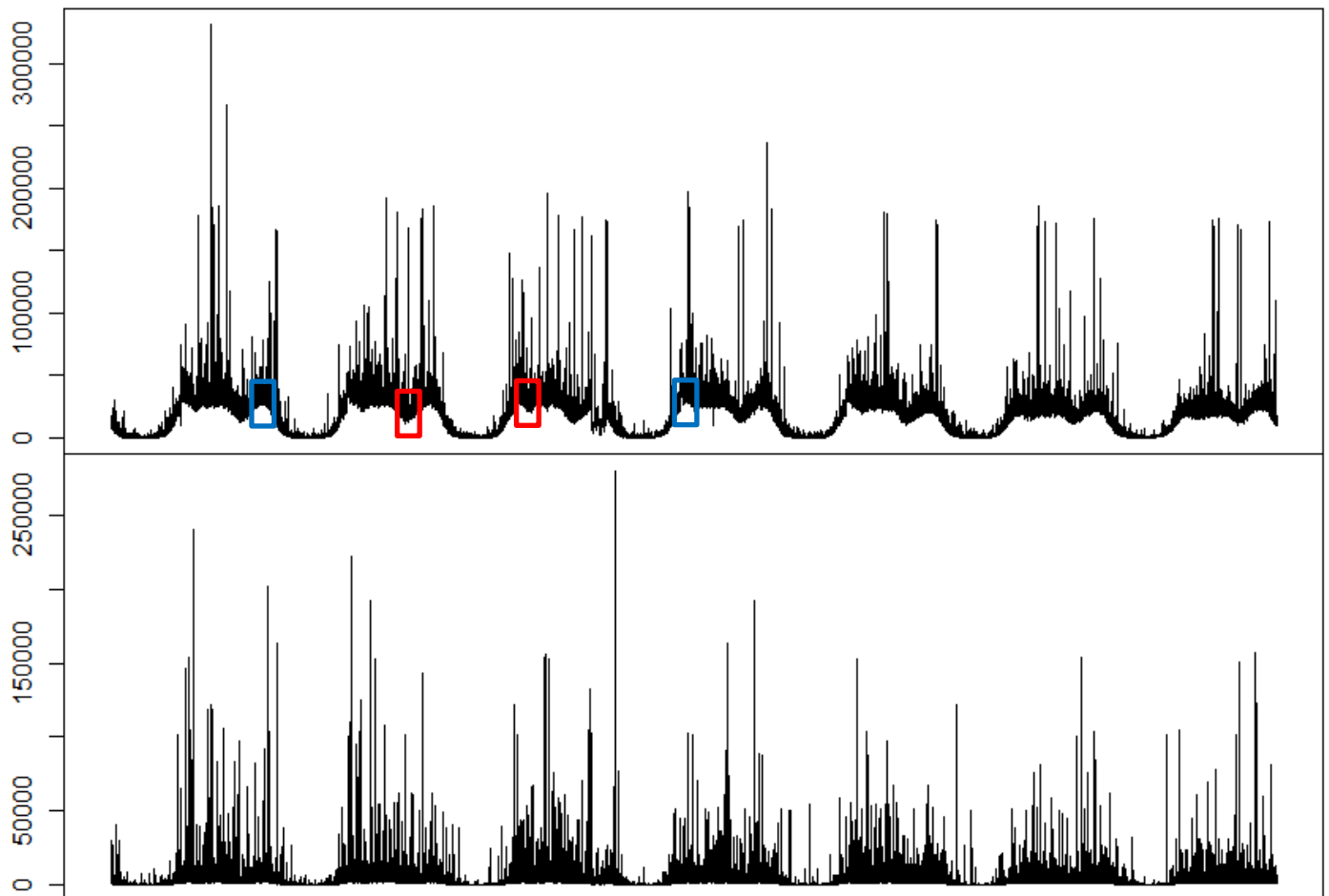


图 1

有趣的是午饭时间(大概在每天的 11 点到 12 点)以及晚饭时间大概在 5 点到 6 点是每天的局部极小值。(是不是说明对中国人来说吃饭最重要?) 如图中红色的方框所示。而午饭与晚饭后的一小时则是交易的黄金时段(图中蓝色框所示)。可能是由于大家午饭后的休息时间经常被用来上网购物的原因。

另外，一周的总体趋势是：在周一周二的时候交易量比较大，而周末的交易量较小。这可能是因为在工作日大家利用上班时间经常会网上购物（在图 1 中差异并没有那么大，图 2 中 trend 项可以看出来）。

利用先前的方法对交易数据做了 STL 分解：

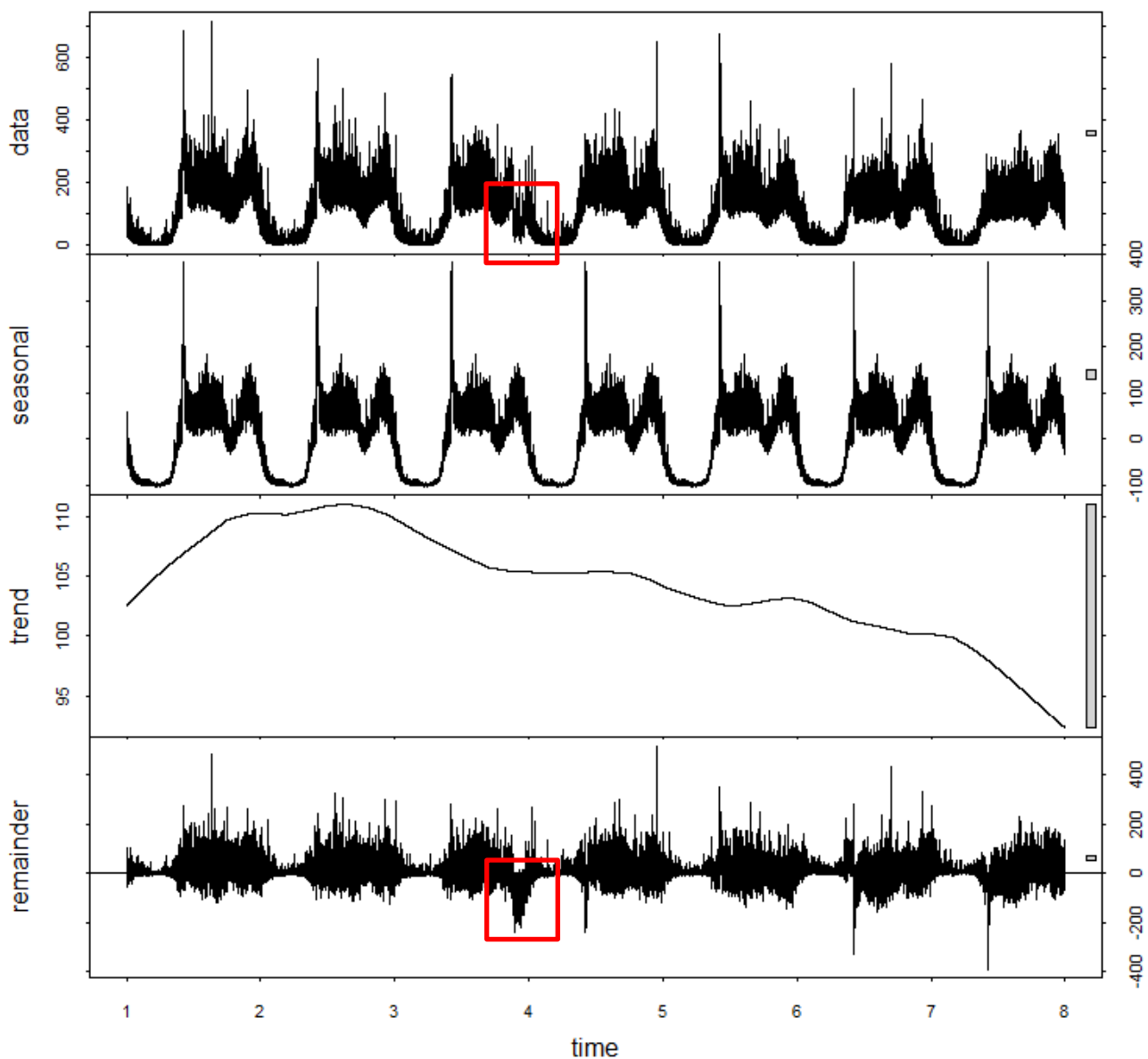


图 2

从趋势项就可以明显反映出交易异常值在周二达到峰值，在周末降到最低点。残余项也基本反映了白天的异常交易值要高于晚上，可能主要是因为白天的总体交易额比较大的缘故。从图中可以看出在星期三（9 月 21 日）的晚饭时间出现了一个极小值。这在残余项中反映了出来，至于为什么会出这样的异常，现在我还没有找到答案。

先前的想法是根据用户选取的结果做层次式的可视化，可是目前的难点在于：程序不可能一次性读入所有数据。但是现在的交易数据全部用文件存储，在读写，排序筛选时显得很麻烦。比如要读取某日 12 点的数据，文件的方式就得从头开始查找一直到定位到 12 点的数据位置。这样显然不实用，而且浪费了大量的时间。因此采用数据库的方式来替代文件读取。将这些数据输入到数据库，一方面对数据的时候会方便很多，另外提高程序运行效率。先采用 `mysql` 数据库存放数据，由于以前没有使用过数据库编程，所以对我来说处理起来有点麻烦。目前的进度，还在配置数据库的编码，使得其能够容纳中文字符串的阶段。另一方面，系统先暂时做了选择时间日期的交互以方便探索数据，如下图所示。

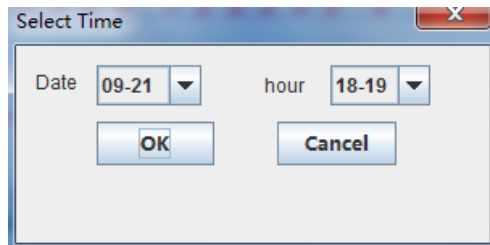


图 3 系统临时的时间选择界面

由图 2 可知周三傍晚出现了交易异常值的极小值，在探索过程中并没有找到原因。总结了一下问题可能有几方面：

1. 我们的系统比较适合寻找异常值的极大值，但是出现交易低谷的情况，就说明没有太多异常交易，因此也无法从音符图中找原因。
2. 从图一中得知，周三晚上的这段时间交易金额处于低谷值，可能仅仅是因为这段时间交易较少，导致异常交易值也比较少。
3. 我们对交易异常值的计算有问题，是否应该按照当时的交易总量，对交易异常值做一下标准化？

下周工作：

1. 了解主动学习在可分析中的应用，探索本系统如何结合主动学习的方法。以此方法避免异常标准选取不恰当的问题。
2. 将文件读写的系统改为利用数据库连接读写数据，提高系统读数据的效率。
3. 改进交易异常值的计算方法。